Bootstrap confidence limits for groundfish trawl survey estimates of mean abundance

Article *in* Canadian Journal of Fisheries and Aquatic Sciences · April 1997

DOI: 10.1139/cjfas-54-3-616

1 author:

Stephen J Smith
Bedford Institute of Oceanography
80 PUBLICATIONS **2,722** CITATIONS

Bootstrap confidence limits for groundfish trawl survey estimates of mean abundance

Stephen J. Smith

Abstract: Trawl surveys using stratified random designs are widely used on the east coast of North America to monitor groundfish populations. Statistical quantities estimated from these surveys are derived via a randomization basis and do not require that a probability model be postulated for the data. However, the large sample properties of these estimates may not be appropriate for the small sample sizes and skewed data characteristic of bottom trawl surveys. In this paper, three bootstrap resampling strategies that incorporate complex sampling designs are used to explore the properties of estimates for small sample situations. A new form for the bias-corrected and accelerated confidence intervals is introduced for stratified random surveys. Simulation results indicate that the bias-corrected and accelerated confidence limits may overcorrect for the trawl survey data and that percentile limits were closer to the expected values. Nonparametric density estimates were used to investigate the effects of unusually large catches of fish on the bootstrap estimates and confidence intervals. Bootstrap variance estimates decreased as increasingly smoother distributions were assumed for the observations in the stratum with the large catch. Lower confidence limits generally increased with increasing smoothness but the upper bound depended upon assumptions about the shape of the distribution.

Résumé : Les relevés au chalut selon des plans d'échantillonnage aléatoire stratifié sont couramment utilisés sur la côte est de l'Amérique du Nord pour surveiller les populations de poisson de fond. Les estimations sont calculées par randomisation et ne nécessitent pas l'établissement d'un modèle probabiliste pour les données. Toutefois, les propriétés de ces estimations, qui correspondent à des échantillons de grande taille, ne conviennent pas nécessairement aux échantillons de petite taille ni aux données asymétriques caractéristiques des relevés au chalut de fond. Dans notre étude, nous avons employé trois méthodes de rééchantillonnage bootstrap intégrant des plans d'échantillonnage complexes pour explorer les propriétés des estimations correspondant à de petits échantillons. Nous présentons une nouvelle formule permettant de calculer les intervalles de confiance après correction du biais et accélération dans les relevés à échantillonnage aléatoire stratifié. Les résultats de simulation indiquent que les limites des intervalles de confiance après correction du biais et accélération peuvent surcorriger les données des relevés au chalut, et que les limites en percentiles étaient plus proches des valeurs prévues. Des estimations non paramétriques de la densité ont servi à étudier les effets de captures anormalement grandes de poisson sur les estimations bootstrap et les intervalles de confiance correspondants. Plus les distributions supposées pour les observations situées dans la strate des fortes captures étaient lisses, plus les estimations bootstrap des variances étaient faibles. Les limites de confiance inférieures montaient généralement en même temps qu'augmentait le degré de lissage, mais la limite supérieure dépendait des hypothèses sur la forme de la distribution.
[Traduit par la Rédaction]

Introduction

Annual bottom trawl surveys provide a major source of fisheries independent information on abundance, species composition, and basic biological data for the groundfish communities of the Atlantic coast of Canada and the United States (Azarovitz 1981; Halliday and Koeller 1981; Pitt 1981). Most of these surveys use a stratified random design with stratum boundaries defined by depth ranges, species-specific distributions, and management areas. The abundance estimates of fish from these surveys are routinely used in the stock assessment of many commercial groundfish species (Smith 1988*a*; Gunderson 1993).

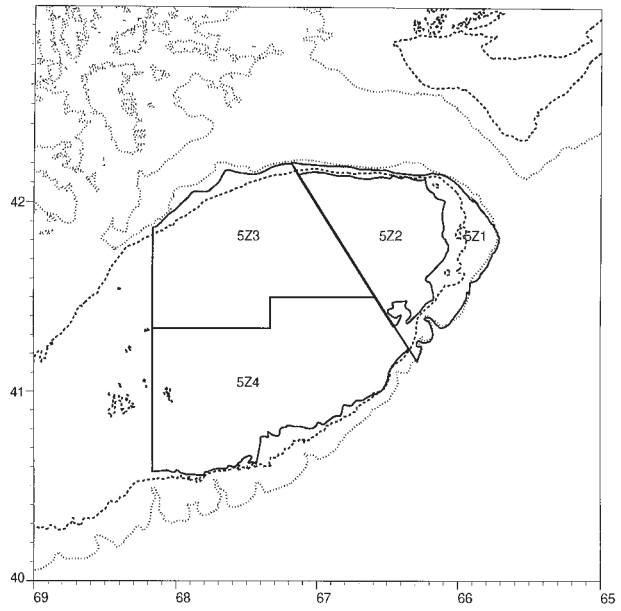
The statistical properties of quantities commonly estimated from stratified random trawl surveys such as mean and total catch are usually estimated using standard methods from sample survey or finite population theory (Smith 1990). Standard error estimates are derived assuming repeated sampling from the finite population (e.g., trawl sites). Confidence intervals are constructed on the basis of the normal distribution, which was shown to be the limiting distribution for the stratified mean (and total) when the central limit theorem was applied to sampling from a finite population (Cochran 1977; Thompson 1992). However, in many cases, small sample sizes within strata and observed fish catches exhibiting skewed frequency distributions result in confidence limits that are not very useful. In particular these confidence limits may have extremely long intervals between them or even negative lower limits on occasion.

Alternatively, a number of statistical models for positive random variables (e.g., Poisson, Δ distribution) have been suggested as possible models for catch per tow (Taylor 1953; Pennington 1983; Smith 1988*b*; McConnaughey and Conquest 1993). The usual advantages given for using such models include (*i*) the provision of more precise estimates of the mean and (*ii*) the possibility (but not necessarily the means) of obtaining confidence intervals that are always positive. However, Smith (1990) has shown how the application of statistical models and their specific estimates to survey data can result in

Received May 24, 1995. Accepted November 21, 1996.
J12927

S.J. Smith. Department of Fisheries and Oceans, P.O. Box 1006, Dartmouth, NS B2Y 4A2, Canada. e-mail: s_smith@bionet.bio.dfo.ca

Fig. 1. Stratification map for Georges Bank surveys (1987–present). Stratum boundaries are based primarily on the average spatial distribution in February–March of Atlantic cod and haddock. The numbers on the map identify the individual strata.

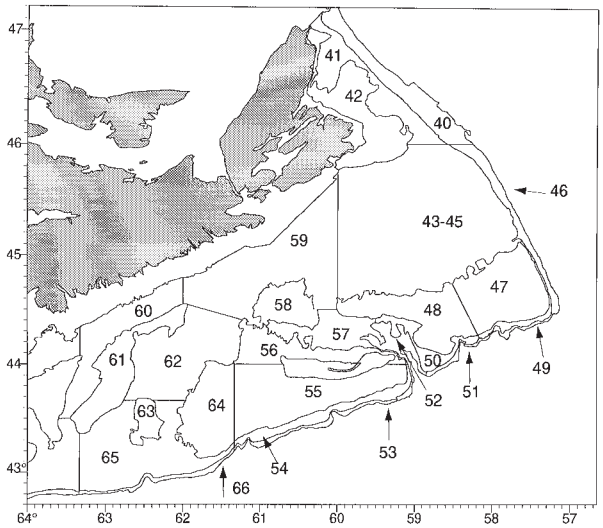


biased estimates of the population means and variances normally estimated in sample survey theory (see also Jolly and Smith 1989; Skinner et al. 1989; Myers and Pepin 1990).

A third approach for modelling the distribution of estimates from surveys is to use bootstrap resampling methods (Efron 1982). Bootstrap methods have been used in a number of fisheries survey applications (e.g., Kimura and Balsinger 1985; Sigler and Fujioka 1988; Robotham and Castillo 1990; Pelletier and Gros 1991; Buckland et al. 1992; Stanley 1992; Smith and Gavaris 1993) as a means of substituting computational power for theoretical analysis in situations requiring complex models or estimates. The bootstrap offers a natural way of modelling survey estimates given that its basis is very similar to that of the randomization basis for finite population theory. Bootstrap confidence intervals do not require a distributional assumption for their construction and thus can be used to evaluate the standard normal theory intervals.

In this paper, I apply the bootstrap methodology to the construction of confidence intervals for the stratified mean number of haddock (*Melanogrammus aeglefinus*) per tow from groundfish trawl surveys of the Scotian Shelf and Georges Bank. The application of the bootstrap methodology to complex survey designs is not straightforward and three published methods for stratified survey designs are compared here. A number of methods exist for estimating bootstrap confidence intervals and three of these methods are compared. A new form for the bias-corrected and accelerated bootstrap confidence interval is presented for stratified survey designs. Nonparametric density estimates are used to evaluate these bootstrap confidence intervals. Finally, nonparametric density estimates are also used to investigate the impact on bootstrap estimates of different hypotheses concerning the frequency distribution of catch data when the occasional extremely large catch of fish is encountered.

Fig. 2. Stratification map for the Scotian Shelf surveys conducted in July (1970–present). Stratum boundaries are based primarily on depth ranges (originally measured in fathoms). The numbers on the map identify the individual strata.



Materials and methods

Trawl survey data

The fish catch data sets used in this study were obtained from the standard groundfish trawl surveys of Georges Bank and the Scotian Shelf conducted by the Marine Fish Division (Bedford Institute of Oceanography, Dartmouth, N.S., and Biological Station, St. Andrews, N.B.) of the Canadian Department of Fisheries and Oceans. These surveys use a stratified random survey design (Cochran 1977; Smith 1988*a*). The Georges Bank strata (Fig. 1) were based on the historical distribution of Atlantic cod (*Gadus morhua*) and haddock during February–March, and in addition, incorporated the Canada–U.S. boundary (Gavaris and Eeckhaute 1990). The eastern Scotian Shelf strata (Fig. 2) were primarily based on depth boundaries of 91, 183, and 366 m (originally 50, 100, and 200 fathoms). Further delineations of the stratum boundaries reflect species and stock distributions (Doubleday 1981; Halliday and Koeller 1981).

The sample unit for the survey is defined as the area over the bottom covered by a trawl of a specific width (12.5 m) towed at 3.5 knots (1 knot = 1.852 km/h) for a distance of 1.75 nautical miles (1 nautical mile = 1.852 km). The positions of these sample units or sets were selected randomly before the cruise for each stratum. The design, operation, and analysis of stratified random trawl surveys on the Scotian Shelf have been discussed by Halliday and Koeller (1981) and Smith (1988*a*). We will need the following definitions for quantities associated with the trawl surveys in any one year: n_h is the number of hauls or sets sampled in stratum h ($h = 1, \ldots, L$); $n = \Sigma_{h=1}^{L} n_h$, the total number of sets sampled in the survey; N_h is the total number of possible sets in stratum h; $N = \Sigma_{h=1}^{L} N_h$, the total number of possible sets in the survey area; $f_h = n_h/N_h$, the sampling fraction in stratum h; $W_h = N_h/N$, the proportion of the area in stratum h; y_{hi} is the number of fish caught in set i and stratum h; $\bar{y}_h = \Sigma_{i=1}^{n_h} y_{hi}/n_h$, the estimated mean abundance in stratum h; and $s_h^2 = \Sigma_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2/(n_h - 1)$, the estimated variance in stratum h.

The stratified mean abundance and its associated variance are estimated as

$$(1) \qquad \bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

and

$$(2)\qquad \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{st}}) = \sum_{h=1}^{L} \frac{N_h}{N^2}(N_h - n_h)\frac{s_h^2}{n_h}$$

respectively.

There are two important aspects of eq. 2 that may not be readily apparent from the formula. In the first place, the variance that is being measured here concerns the effectiveness of estimating the mean catch over all N_h units, $\bar{Y}_h$, with the sample mean, that is, the expected value over all strata of $(\bar{y}_h - \bar{Y}_h)^2$ for all distinct samples of size n_h repeatedly chosen from the N_h possible sample units within each stratum. This formulation is unaffected by any spatial structure that may exist for the species being captured, although temporal stability of the fish distribution is implicitly assumed. Secondly, this estimate gives the variance of the sample mean as a predictor of what may be expected to be caught in those sites where trawls were not made (see, for example, Smith 1990). The more sample units that are observed, the better the sample mean will be as a predictor of fish catch.

Parametric confidence intervals for the stratified mean are constructed by assuming that under repeated sampling the $\bar{y}_{st}$ have a normal or Student's t distribution (Cochran 1977, pp. 95–96). The effective degrees of freedom for the Student's t are estimated as

$$\mathrm{df}_e = \frac{\left(\sum_{h=1}^{L} g_h s_h^2\right)^2}{\sum_{h=1}^{L} \frac{g_h^2 s_h^4}{n_h - 1}}$$

where $g_h = N_h(N_h - n_h)/n_h$. This method is valid even if the variances differ between strata. However, the method does require that the y_{hi} are normally distributed so that the individual terms of the sum in the denominator are estimates of the variance of s_h^2.

Bootstrap methods for stratified random designs

The bootstrap resampling method has been successfully used to provide standard error estimates and nonparametric confidence intervals for both simple linear statistics and nonsmooth functions of data (Efron 1982). The basic idea of the bootstrap is to treat the original sample of size n as the target population and the original estimated statistic (e.g., mean, ratio) as the population parameter to be estimated. Repeated sampling with replacement of size n from the original data set is used to create a large number of new pseudosamples. Estimates of the parameter of interest are made for each of the pseudosamples and the empirical distribution of these resultant estimates is used to characterize the distribution of the original statistic. Bias is evaluated with respect to the difference between the average of all of the bootstrap estimates and the original estimate. In cases where it is known that the original statistic and its variance are unbiased, the bootstrap estimate and its variance estimate must also be so.

The bootstrap method was originally introduced for independently and identically distributed cases where all sample units had an equal probability of being chosen. In complex survey designs, such as stratified random designs, the sample units have the same probability of being chosen within any one stratum but different strata may have very different sampling intensities. Following are three different approaches of extending the bootstrap method to sample survey estimates from complex survey designs.

Naïve bootstrap

The simplest (and hence the most naïve) approach to applying the bootstrap to samples from a stratified random design is to resample the observations independently within each stratum, that is,

1. Within stratum h take a simple random sample of the y_{hi} of size n_h with replacement and calculate $\bar{y}_h^*$. Repeat independently for all L strata.

2. Calculate the stratified mean as $\bar{y}_{\mathrm{st},b} = \sum W_h \bar{y}_h^*$.

3. Repeat steps 1 and 2 a total of B times to get $\bar{y}_{\mathrm{st},1}, \ldots, \bar{y}_{\mathrm{st},B}$.

The bootstrap estimate of the stratified mean is calculated as

$$(3)\qquad \bar{y}_{\mathrm{st}}^* = \sum_{b=1}^{B} \bar{y}_{\mathrm{st},b}/B$$

with variance given by

$$(4)\qquad \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{st}}^*) = \frac{1}{B-1}\sum_{b=1}^{B}(\bar{y}_{\mathrm{st},b}^* - \bar{y}_{\mathrm{st}}^*)^2$$

Given that the observed stratified mean (eq. 1) and variance (eq. 2) are unbiased estimates we would expect the expected values of $\bar{y}_{\mathrm{st}}^*$ and $\mathrm{Var}(\bar{y}_{\mathrm{st}}^*)$ to equal their respective observed values. While the bootstrap estimate of the mean is an unbiased estimate of the stratified mean, the expectation of the bootstrap variance has been shown to be equal to (Efron 1982; Rao and Wu 1988)

$$(5)\qquad E(\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{st}}^*)) = \sum_{h=1}^{L} W_h^2 \left(\frac{n_h - 1}{n_h}\right)\frac{s_h^2}{n_h}$$

and therefore the naïve bootstrap variance is an inconsistent estimate of the variance in eq. 2. While a correction could be derived to obtain a consistent estimate of the variance, this would be missing the point, because the purpose here is to generate a population of bootstrap observations with the expected mean and variance.

Rescaling bootstrap

Rao and Wu (1988) proposed rescaling the resampled values of resample size m_h ($m_h \le n_h$) so that the resulting bootstrap variance estimate would be equal to its expected value in eq. 2. Determination of the appropriate resample sizes is the key to finding the most consistent estimates of the variance. The method of Rao and Wu (1988) proceeds as follows:

1. Take a simple random sample of size m_h ($\le n_h$) with replacement from the given sample in each stratum h ($y_{hi}^{\bullet}$, $i = 1, \ldots, m_h$). This is done independently for each stratum. Calculate

$$\tilde{y}_{hi} = \bar{y}_h + \left(\frac{m_h(1 - f_h)}{(n_h - 1)}\right)^{1/2}(y_{hi}^{\bullet} - \bar{y}_h)$$

$$\tilde{y}_h = m_h^{-1}\sum_{i=1}^{m_h}\tilde{y}_{hi}$$

$$\tilde{y}_{\mathrm{st}} = \sum_{h=1}^{L} W_h \tilde{y}_h$$

2. Replicate step 1 B times to obtain $\tilde{y}_{\mathrm{st},1}, \ldots, \tilde{y}_{\mathrm{st},B}$.

The bootstrap estimates of the stratified mean and variance are calculated in exactly the same way as for the naïve bootstrap by substituting $\tilde{y}_{\mathrm{st},b}$ for $\bar{y}_{\mathrm{st},b}^*$ in eqs. 3 and 4.

Rao and Wu (1988) suggested setting $m_h = n_h - 3$, on the basis of comparing the bootstrap third moment $E_*(\tilde{y}_{\mathrm{st},b} - \tilde{y}_{\mathrm{st}})^3$ with the unbiased estimate of the third moment of $\bar{y}_{\mathrm{st}}$. In a recent study, Kovar et al. (1988) compared the merits of setting $m_h = n_h - 3$ to $m_h = n_h - 1$ and found that their results favoured the latter over the former. Smith and Gavaris (1993) reported similar results for a limited example. However, we will only be able to compare $m_h = n_h - 3$ with $m_h = n_h - 1$ for the Georges Bank example because a number of the n_h in the Scotian Shelf survey were less than or equal to 3.

While the rescaling method has been found to provide more unbiased estimates of the variance than the naïve method, it has the disadvantage of being a more computer-intensive method.

Table 1. Summary of quantities used in calculating the stratified mean and variance for the number of haddock caught in the 1989 Georges Bank survey.

Stratum (h)	n_h	W_h	$\bar{y}_h$	s_h
5Z1	18	0.1622	27.82	23.90
5Z2	26	0.2555	85.94	120.99
5Z3	8	0.3069	2.19	5.78
5Z4	6	0.2755	1.05	1.74

Note: The sample size, mean, and standard deviation for each stratum are given by n_h, $\bar{y}_h$, and s_h, respectively. The proportion of the total survey area in each stratum is given in the column headed by W_h.

Mirror-match bootstrap

Sitter (1992*a*) designed the mirror-match bootstrap to imitate the original within-stratum sampling scheme and provide unbiased estimates of the stratified variance.

1. Take a simple random sample without replacement of size m_h ($\leq n_h$) from stratum h to obtain a new set of observations $y'_{h1}, \ldots, y'_{h,m_h}$.

2. Repeat step 1 k_h times independently, replacing the resamples of size m_h each time to get k_h groups of observations.

3. Calculate the mean for stratum h as $\bar{y}'_h = \Sigma_{j=1}^{k_h} \Sigma_{l=1}^{m_h} y'_{hjl}/(k_h m_h)$, where l and j refer to the observations chosen in the without- and with-replacement steps.

4. Repeat steps 1 and 2 for each stratum and calculate the stratified mean as per eq. 1, replacing $\bar{y}_h$ with $\bar{y}'_h$.

5. Replicate step 1 B times to obtain $\bar{y}'_{st,1}, \ldots, \bar{y}'_{st,B}$.

As in the previous case, the bootstrap estimates of the stratified mean and variance are calculated in exactly the same way as for the naïve bootstrap by substituting $\bar{y}'_{st,b}$ for $\bar{y}^*_{st,b}$ in eqs. 3 and 4.

In the first step m_h is chosen such that the sampling fraction m_h/n_h is as close as possible to that for the original sample n_h/N_h. This step is repeated k_h times, where $n_h = k_h m_h$ so that the final sample size is the same as the original sample size. However, if $n_h/N_h \leq 1/n_h$ then the replication factor in step 2 is calculated as $k_h = n_h(1 - m_h/n_h)/m_h(1 - n_h/N_h)$. The sample size in step 1 is obtained as $m_h = n_h(1 - m_h/n_h)/(1 - n_h/N_h)$. If either k_h or m_h is noninteger, a randomization between bracketing integers is used.

The sampling fractions n_h/N_h for the data in this study are all <0.0003 whereas the smallest $1/n_h$ is equal to 0.038 and therefore m_h defaults to 1.0. The replication factor k_h will be a noninteger bracketed by $n_h - 1$ and n_h. Following Sitter (1992*a*) an integer value k'_h is chosen by using a Bernoulli random number generator,

$$k'_h = \begin{cases} n_h - 1 & \text{when random number} = 1 \\ n_h & \text{when random number} = 0 \end{cases}$$

where the probability of obtaining a random number = 1 is

$$P(x=1) = \frac{\left(\frac{(1-f_h)}{(n_h-1)} - \frac{1}{n_h}\right)}{\left(\frac{1}{n_h-1} - \frac{1}{n_h}\right)}$$

This special case of the mirror-match method (i.e., $m_h = 1$) is equivalent to the bootstrap with replacement (BWR) method of McCarthy and Snowden (cited in Sitter 1992*a*). This method is also similar to the naïve method with the addition of a randomization step for choosing either n_h or $n_h - 1$ resamples within each stratum.

Confidence intervals from bootstrap resampling

There are a number of ways of computing percentiles for confidence intervals from bootstrap estimates (see Efron and Tibshirani 1993). The three considered here, the percentile method (PC), the bias-corrected method (BC), and the bias-corrected and accelerated method (BC_a), are the more commonly used methods. The PC method assumes that the frequency distribution of the bootstrap estimates fully describes the distribution function of some estimate $\hat{\theta}$, $G(s) = \text{Prob}(\hat{\theta} \leq s)$. That is,

$$\hat{G}(s) = \sum_{i=1}^{B} I_i / B$$

where

$$I_i = \begin{cases} 1, & \text{if } \hat{\theta}^*_i \leq s; \\ 0, & \text{otherwise.} \end{cases}$$

and $\hat{\theta}^*_i$ denotes the ith bootstrap estimate of θ (e.g., stratified mean) and B as before ($i = 1, \ldots, B$) denotes the number of bootstrap replications. Upper and lower α confidence intervals are calculated as $\hat{G}^{-1}(1 - \alpha/2)$, $\hat{G}^{-1}(\alpha/2)$, respectively.

The BC method (Efron 1981) introduces a correction to the PC method to account for differences between $\hat{\theta}$ and the median of the frequency distribution. The α upper and lower confidence intervals for this bias-corrected method are obtained as $\hat{G}^{-1}(\Phi(z^{(1-\alpha/2)} + 2z_0))$, $\hat{G}^{-1}(\Phi(z^{(\alpha/2)} + 2z_0))$, respectively; where Φ is the standard normal distribution function, z^t is the tth percentile of the standard normal distribution, and

$$z_0 = \Phi^{-1}\left(\hat{G}^{-1}\left(\frac{\#(\hat{\theta}^*_i < \hat{\theta})}{B}\right)\right)$$

(where #( ) refers to a count of how many times the condition within the parentheses is true). The term z_0 will be equal to zero when the bootstrap estimate and the median of the $\hat{\theta}^*_i$ are equal; the bias-corrected and percentile methods are equivalent in this case.

Finally, the standard normal approximation often used in constructing confidence intervals assumes that the mean is independent of the variance. However, in many cases and certainly for trawl survey catch data this assumption does not hold. A further correction factor, a, referred to as the acceleration is introduced as a measure of the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ measured on a normalized scale. The α upper and lower confidence intervals for the BC_a method are given as

$$\hat{G}^{-1}\left(\Phi\left(z_0 + \frac{z_0 + z^{(1-\alpha/2)}}{1 - \hat{a}(z_0 + z^{(1-\alpha/2)})}\right)\right)$$

and

$$\hat{G}^{-1}\left(\Phi\left(z_0 + \frac{z_0 + z^{(\alpha/2)}}{1 - \hat{a}(z_0 + z^{(\alpha/2)})}\right)\right),$$

respectively.

Note that setting both z_0 and $\hat{a}$ to zero would result in the formula for the percentile confidence intervals. The acceleration is estimated here using the jackknife-based estimate (Efron and Tibshirani 1993, p. 186) modified for stratified random surveys (see Appendix).

Results

Distribution of survey data

The data used in this paper are from the 1989 winter groundfish survey on the eastern portion of Georges Bank and the 1988 summer groundfish survey of the Scotian Shelf. The haddock catches from each survey are summarized in Tables 1 and 2. The abundance indices for the haddock stock on Georges Bank were estimated from catches in strata 5Z1–5Z4 (Fig. 1). The four strata were fairly similar in size (W_h) and enjoyed a higher sampling intensity per stratum than many of the surveys in the Northwest Atlantic. For the eastern Scotian Shelf haddock

stock, abundance indices were estimated from catches in strata 40–66 (Fig. 2). This survey was more characteristic of the usual situation in the Northwest Atlantic (Doubleday 1981) by having many strata of highly variable size, with the number of sets per stratum ranging from 2 to 8 in most years.

The frequency distributions for both data sets exhibited highly skewed distributions with large proportions of zeros (greater than 40% of the observations) and long right-hand tails (Fig. 3). The empirical cumulative distribution functions were estimated using the standard estimator for stratified random designs given in Chambers and Dunstan (1986). Up to a point the distributions were similar for the two areas; however, there was one extremely large catch of 5496 haddock in the eastern Scotian Shelf survey. In fact, this was the largest catch of haddock in the history of this survey.

Bootstrap estimate of the variance of the stratified mean

Before proceeding to investigate the bootstrap distributions of the stratified mean, the variance estimates (eq. 4) from the three bootstrap methods, naïve, rescale, and BWR, were compared with the estimate from the usual stratified variance of the mean (eq. 2). The percent deviation of the various bootstrap estimates from the stratified variance estimate is presented as a function of the number of replications (B) in Fig. 4 for the Georges Bank data set and in Fig. 5 for the eastern Scotian Shelf. As advertised, the naïve method resulted in a biased estimate of the variance. All of the other methods gave reasonable estimates of the variance, with the BWR method having the minimum deviation for the Scotian Shelf data set and the rescale method with $m_h = n_h - 1$ performing better for the Georges Bank data set. However, differences between the rescale methods and the BWR method were quite minimal. The naïve method will be retained for the rest of this study for comparative purposes even though the variance estimates were biased.

Another aspect of the comparison in Figs. 4 and 5 is an appreciation for the number of bootstrap replications required to get stable variance estimates. Reference levels for bootstrap resample sizes have been given as 20–50 for estimating standard errors (Efron and Tibshirani 1993, p. 273); however, these results indicate that much higher levels are required when dealing with complex survey designs. Variance estimates for the rescale and BWR methods converged much more quickly to the true value for the Georges Bank data set ($B \approx 300$) than they did for the Scotian Shelf data ($B \approx 750$), possibly a reflection of the higher sampling intensity in the former data set.

Bootstrap distribution of stratified mean

The empirical distribution curves for 1000 bootstrap estimates of the stratified mean for each of the four bootstrap methods for the Georges Bank data (Fig. 6) were quite smooth and there appeared to be very little difference between the different bootstrap methods. On the other hand, the empirical cumulative distributions for the eastern Scotian Shelf data plotted in Fig. 7 showed a number of differences. The distributions for the rescale and BWR methods were quite similar whereas the naïve method deviated from the other two curves at the 35th percentile and higher. All of the curves showed a number of discontinuities in the curves where few or no bootstrap estimates of the stratified mean were obtained. These discontinuities were the result of a problem common to trawl surveys: one very large catch of fish in the survey. Recall that in Fig. 3, the maximum observation for the 4VW haddock survey was 5496 fish, which was a considerable catch compared with the next largest catch of 309 haddock. This next largest catch also occurred in the same stratum (56) as the maximum catch. This large catch of 5496 haddock was quite influential in that by itself it accounted for 49% of the stratified mean. In contrast, the largest catch in the Georges Bank survey (425) only accounted for 15% of its stratified mean.

Table 2. Summary of quantities used in calculating the stratified mean and variance for the number of haddock caught in the 1988 eastern Scotian Shelf survey.

Stratum (h)	n_h	W_h	$\bar{y}_h$	s_h
40	6	0.0294	0.34	0.84
41	4	0.0318	7.40	5.50
42	7	0.0457	0.77	0.86
43	4	0.0419	0.26	0.51
44	4	0.1247	1.29	2.57
45	4	0.0325	0.00	0.00
46	3	0.0156	0.00	0.00
47	6	0.0513	24.67	45.03
48	5	0.0460	0.00	0.00
49	2	0.0046	17.73	11.32
50	3	0.0122	75.72	85.38
51	2	0.0047	2.06	2.91
52	2	0.0110	55.70	77.40
53	2	0.0082	0.00	0.00
54	2	0.0159	45.80	44.40
55	7	0.0674	94.90	59.60
56	6	0.0303	985.90	2212.80
57	2	0.0258	18.00	25.50
58	3	0.0209	77.90	118.90
59	6	0.1000	24.90	48.80
60	3	0.0427	18.60	17.20
61	2	0.0367	0.00	0.00
62	4	0.0672	2.70	4.80
63	2	0.0096	89.70	26.30
64	5	0.0412	109.30	110.50
65	8	0.0757	63.90	65.60
66	2	0.0072	1.50	2.20

Note: The sample size, mean, and standard deviation for each stratum are given by n_h, $\bar{y}_h$, and s_h, respectively. The proportion of the total survey area in each stratum is given in the column headed by W_h.

The discontinuities in Fig. 7 separate segments of the curve, which going from the bottom to the top of the figure give the distribution of the stratified mean when the catch of 5496 haddock was not chosen or chosen once, twice, or three or more times for the bootstrap sample in its stratum. The fact that this one catch would occur more than once in a bootstrap sample is not too surprising given that there were only six sets taken in stratum 56 and resampling was done with replacement.

All of the salient quantities being estimated for both data sets are presented in Table 3. The estimates of the stratified mean for all methods were within less than 1% of their respective original observed values. The variance estimates in Table 3 verify that the expected values of the naïve variance estimate were well predicted by eq. 5 and that all of the other bootstrap estimates were quite close to their respective original observed variance estimates.

Confidence intervals calculated using the methods given in

Fig. 3. Empirical cumulative distribution plots for numbers of haddock caught from each tow in eastern Scotian Shelf (1988) and Georges Bank bottom trawl surveys (1989). Note that the maximum observation for the eastern Scotian Shelf survey was 5496 haddock.

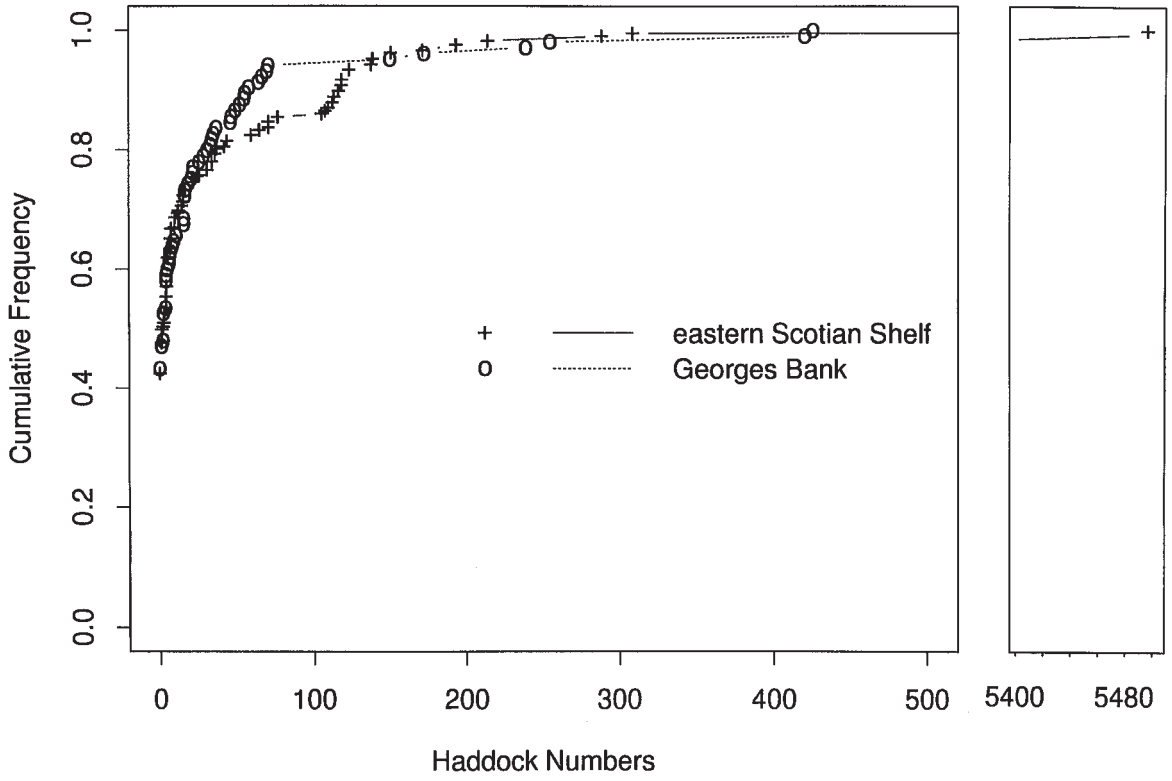


Fig. 4. Comparison of the performance of the variance estimates from each of four bootstrap methods presented here as a function of replication number. Performance was measured as the percent deviation of the bootstrap estimate from the original variance of the stratified estimate. Variance estimates are for the bootstrap stratified mean number of haddock per tow from the Georges Bank survey (1989).

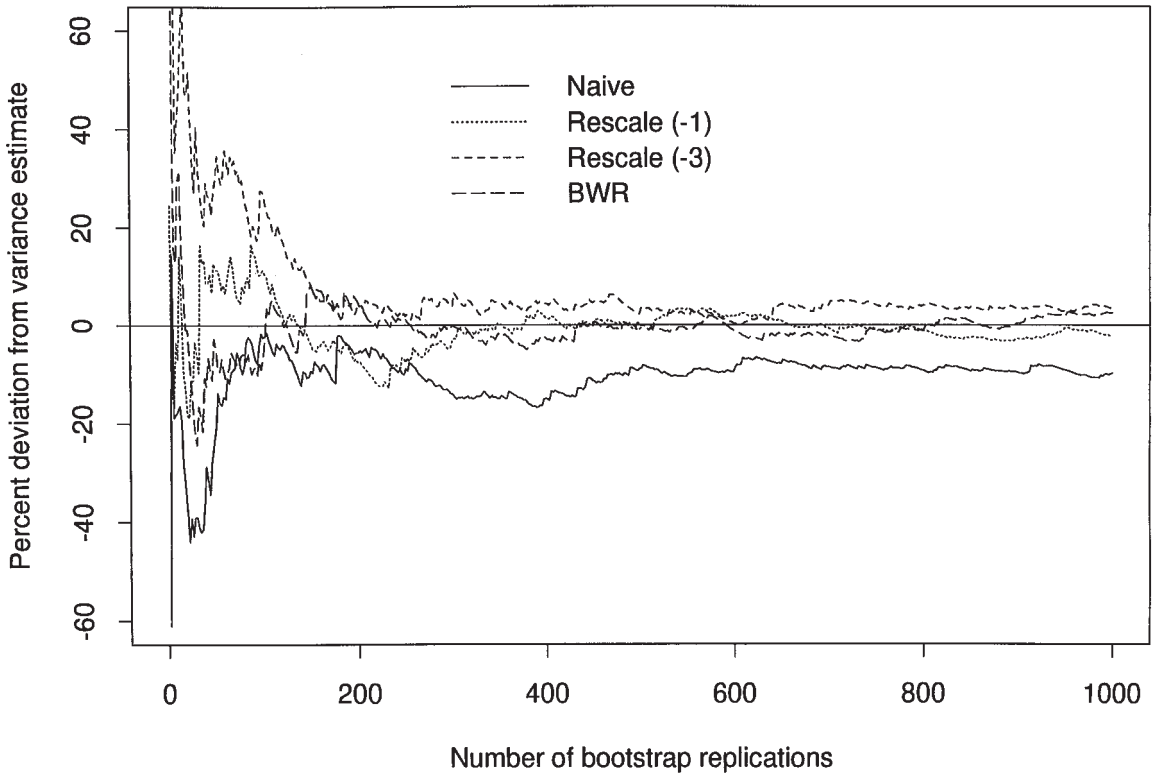


Fig. 5. Comparison of the performance of the variance estimates from each of three bootstrap methods presented here as a function of replication number. Performance was measured as the percent deviation of the bootstrap estimate from the original variance of the stratified estimate. Variance estimates are for the bootstrap stratified mean number of haddock per tow from the eastern Scotian Shelf survey (1988).

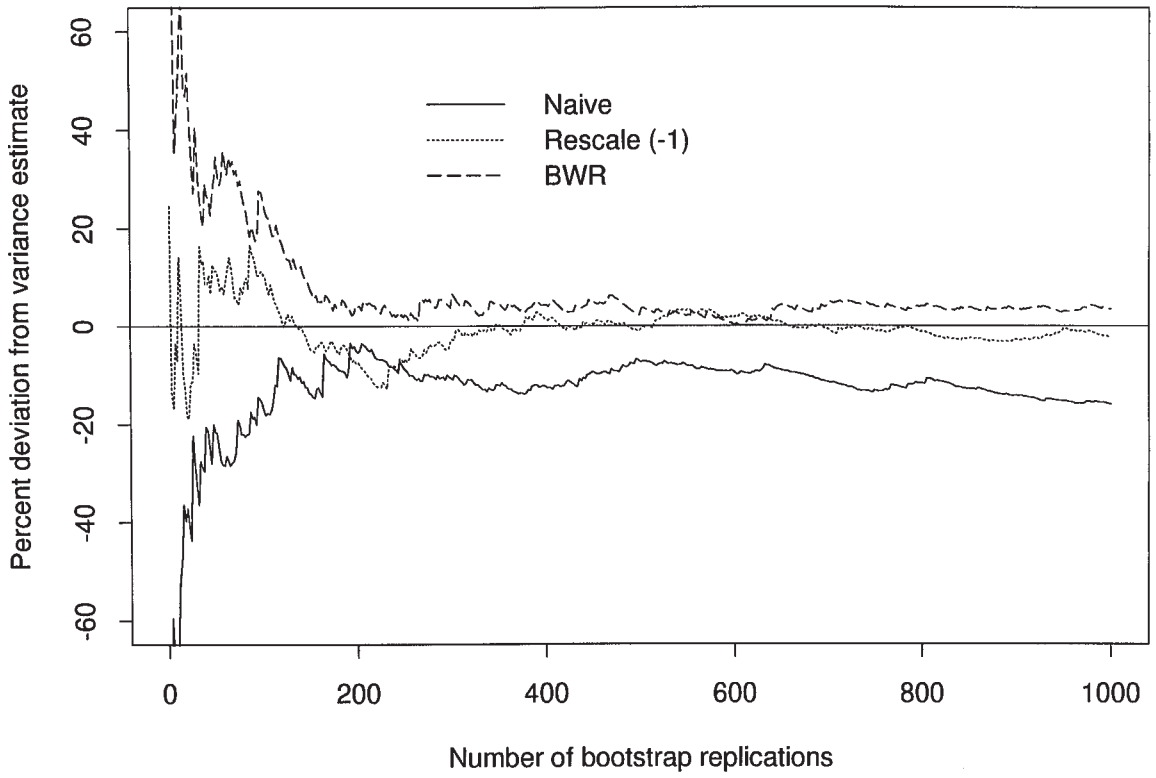


Fig. 6. Empirical cumulative distributions of the stratified mean number of haddock per tow from the Georges Bank survey (1989) estimated by four bootstrap methods.

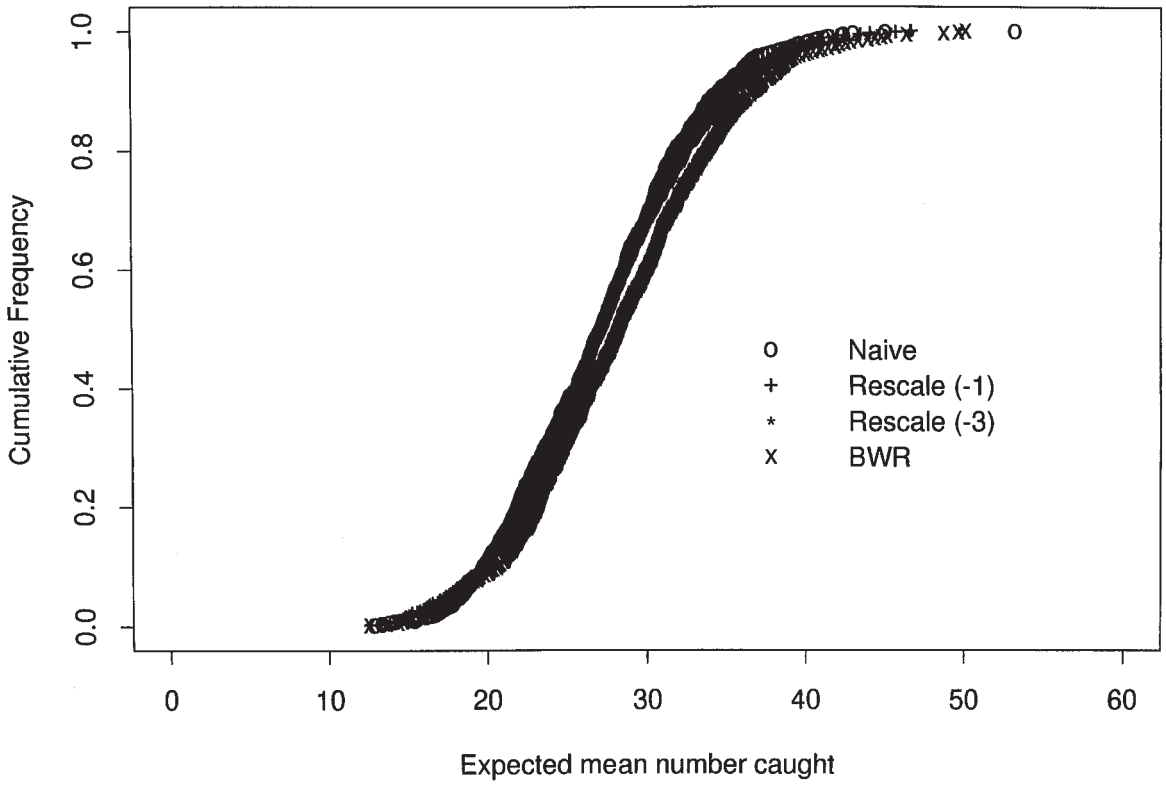
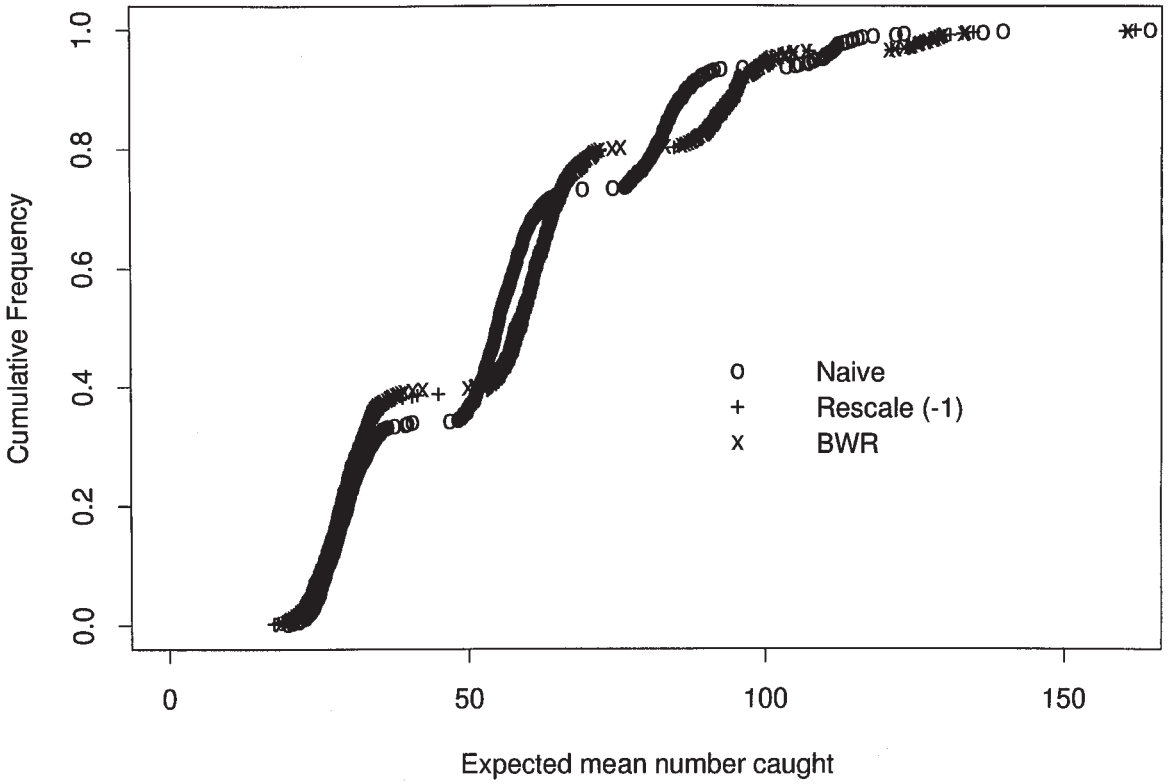


Fig. 7. Empirical cumulative distributions of the stratified mean number of haddock per tow from the eastern Scotian Shelf survey (1988) estimated by three bootstrap methods.



this paper are presented in Table 3 along with information on their length and shape. Shape is calculated as the natural log of the ratio of the upper limit minus the median to the median minus the lower limit (Efron 1992). Therefore, a confidence interval that is symmetric around the median will have a shape measure of zero whereas a shape greater than zero indicates distributions skewed to the right.

While the Student's *t* confidence intervals for the Georges Bank data set appear to be reasonable, the upper and lower bounds for the eastern Scotian Shelf haddock illustrate the problems with using the normal approximation. That is, the intervals can be quite long and the lower limit can be less than zero. The longer interval for the eastern Scotian Shelf survey reflects the small sample sizes within strata and the large variance from stratum 56 where the catch of 5496 haddock was made.

Citing general findings in the literature, Cochran (1977) suggested that $1 - \alpha$ confidence intervals for the mean based on the normal assumption will behave as follows when the original data have a skewed distribution. First, the area covered between the upper and lower limits will be less than $1 - \alpha$. Additionally, the probability of the mean being less than the lower limit will be less than $\alpha/2$ whereas the probability of it being greater than the upper limit will be greater than $\alpha/2$. If the bootstrap limits are to be an improvement then their respective upper and lower limits should be greater than those from the Student *t* distribution.

All of the bootstrap intervals resulted in positive lower limits that were all higher than those given by the Student *t* method. For the case of the eastern Scotian Shelf, these lower limits were substantially higher than the negative lower limit from the Student *t*. However, the bootstrap upper limits were not always greater than the respective Student *t* upper limits. Considering only the unbiased bootstrap methods, bootstrap upper and lower limits increased for each confidence interval method progressing from rescale with $m_h = n_h - 1$, $m_h = n_h - 3$ to the BWR method. Within any one bootstrap method, BC limits were greater or less than their respective PC limits depending upon whether or not the median was less or greater than the bootstrap estimate. In all cases the BC_a upper and lower limits were greater those of the other bootstrap methods and the Student *t* method. Within each bootstrap method the BC_a intervals were the longest and had the most skew, that is, the highest shape measure. The estimated acceleration was less for the Georges Bank data ($a = 0.06$) than for the eastern Scotian Shelf ($a = 0.12$), possibly reflecting the lower variability of the former relative to the latter.

Given that the naïve bootstrap underestimated the variance of the stratified mean it was expected that the corresponding confidence intervals would then be shorter than those for the other bootstrap methods. While this was true for all of the bootstrap methods and confidence interval methods for the Georges Bank data it was only true for the PC limits for the eastern Scotian Shelf. The larger correction for the median (acceleration) for the naïve bootstrap in this latter data set appeared to have compensated for the underestimation of the variance.

Coverage of confidence intervals

Given the observed data only, it is difficult to say how good the confidence intervals in Table 3 really were because there is no theoretical statistical distribution for survey catches or mean

Table 3. Summary of results from calculating the stratified mean, median of the bootstrap distribution, variance, and 95% confidence intervals for the number of haddock caught in the 1989 Georges Bank survey and the 1988 eastern Scotian Shelf survey.

				95% confidence interval				
Method	Mean	Median	Variance	Type	Lower	Upper	Length	Shape
Georges Bank survey								
Original	27.43	*	38.00	St	14.77	40.08	25.31	0.00
Naïve bootstrap	27.16	27.17	34.27	PC	17.04	39.48	22.44	0.19
				BC	17.03	39.46	22.43	0.19
				BC_a	17.88	41.02	23.14	0.40
Rescale bootstrap								
$m_h = n_h - 1$	27.36	27.20	37.08	PC	15.70	39.28	23.58	0.05
				BC	16.15	39.70	23.55	0.06
				BC_a	17.26	41.49	24.23	0.30
$m_h = n_h - 3$	27.50	27.04	39.31	PC	16.49	40.02	23.53	0.21
				BC	17.05	40.98	23.93	0.20
				BC_a	18.08	44.09	26.01	0.52
BWR bootstrap	27.60	27.07	38.95	PC	16.82	40.93	24.11	0.30
				BC	17.56	41.67	24.11	0.26
				BC_a	18.55	43.97	25.42	0.52
Eastern Scotian Shelf survey								
Original	56.15	*	769.1	St	–14.20	126.50	140.70	0.00
Naïve bootstrap	56.19	54.89	646.4	PC	23.69	112.20	88.51	0.61
				BC	24.64	115.71	91.08	0.56
				BC_a	25.88	139.62	113.70	0.94
Rescale bootstrap								
$m_h = n_h - 1$	56.54	58.93	768.6	PC	21.94	124.82	102.88	0.58
				BC	20.35	99.84	79.49	1.40
				BC_a	22.26	127.12	104.90	1.87
BWR bootstrap	56.41	58.21	769.7	PC	22.36	125.46	103.10	0.63
				BC	20.96	102.79	81.84	1.09
				BC_a	22.83	128.28	105.40	1.45

Note: The confidence intervals for the original method were calculated assuming a Student's *t* distribution (St). Confidence intervals for the bootstrap mean were calculated using the percentile (PC), bias-corrected (BC), and bias-corrected accelerated (BC_a) methods. The length column refers to the length of the confidence intervals. The shape column refers to a measure of symmetry (symmetric: shape = 0) given by Efron (1992). Note that the expected values for the variances of the naïve bootstraps (eq. 5) for Georges Bank and the eastern Scotian Shelf were 36.59 and 640.30, respectively.

*By definition equivalent to the mean for the Student *t* distribution.

survey catches within a stratum. While the lower limits from the bootstrap methods tended to move in the right direction relative to the Student *t*, the bootstrap upper limits were not always greater than those from the normal approximation. Also, the limits themselves do not help evaluate whether or not the probabilities of being less than the lower or greater than the upper limit were actually 0.025 with 0.95 between the limits.

In lieu of not having a statistical model for these data, an alternative approach was taken here. Statistical distributions were fitted to the catches within each stratum of the Georges Bank data using nonparametric density functions (Silverman 1986; Scott 1992). In this approach, discrete kernel estimates were used (Rajagopalan and Lall 1995) with different bandwidths for each stratum. The theoretical observations ranged from zero to the maximum catch plus the bandwidth. The fitted distributions were assumed to represent a predictive distribution of what catches could be expected in each stratum (Jones and Bradbury 1993). Random observations were generated from the cumulative mass function for each stratum using the same number of samples per stratum as in Table 1. A total of 10 000 replications of the stratified mean were generated. The 0.025 and 0.975 limits were 16.64 and 40.39 (shape = 0.25), respectively, assuming that these 10 000 replications represent the population distribution for the stratified mean. The lower and upper limits for each method in Table 3 (Georges Bank data) were compared with the population distribution of the stratified mean. The actual proportions of observations in the population less than the lower and greater than the upper are given in Table 4. In addition, the proportions of observations between each upper and lower limit are also given in this table.

The Student *t* interval was wider than expected and the asymmetry in the actual distribution of the stratified mean shows that both upper and lower limits were too low, as was predicted for the normal approximation. The higher bootstrap lower limits result in the associated probabilities being closer to the expected value of 0.025 than for the Student *t* limit. Overall, the PC limits for the rescale bootstrap with $m_h = n_h - 3$ and the BWR bootstrap are closest to having 0.025 in the tails of all of the limits for all of the methods. For the Georges Bank data, the adjustments for the median and acceleration in the BC and BC_a limits appeared to have increased the confidence limits higher than they needed to be for the α level used here.

The lower than expected coverage for the naïve bootstrap reflects the previously noted underestimation of the stratified variance. The coverage was closest to 0.95 for the PC limits

Table 4. Comparison of Student *t* and bootstrap confidence intervals with population confidence intervals constructed from 10 000 replications generated from nonparametric mass functions fitted to the Georges Bank data.

Method	Type	Lower	Upper	Coverage
Original	St	0.008	0.028	0.964
Naïve bootstrap	PC	0.030	0.033	0.937
	BC	0.029	0.034	0.937
	BC_a	0.045	0.020	0.934
Rescale bootstrap				
$m_h = n_h - 1$	PC	0.014	0.036	0.951
	BC	0.018	0.031	0.952
	BC_a	0.032	0.018	0.950
$m_h = n_h - 3$	PC	0.022	0.029	0.950
	BC	0.030	0.021	0.949
	BC_a	0.049	0.009	0.942
BWR bootstrap	PC	0.026	0.022	0.952
	BC	0.038	0.017	0.945
	BC_a	0.060	0.010	0.930

Note: Entries in lower and upper columns refer to the probabilities from the population curve relative to the lower and upper bound from each of the methods used herein. Coverage refers to the probability between the upper and lower limits. The expected probability is 0.025 each for upper and lower limits and the expected coverage is 0.95.

for the rescale and BWR bootstraps with the corrections by the BC_a method being far too severe, particularly for the BWR bootstrap interval.

These results are only directly applicable to the discrete nonparametric density estimate for Georges Bank data. There were too few observations within many of the strata of the Scotian Shelf survey to fit nonparametric densities.

Large catches

The normal approximation and the three types of bootstrap confidence limits were all affected by the large catch of 5496 haddock in the eastern Scotian Shelf survey. Such a large catch appears to be atypical when compared with the other observations in the same survey. From the finite population prediction point of view (see Smith 1990), the estimates of the mean and variance for stratum 56 assume that approximately one sixth of the area was represented by this large catch, an assumption we may believe to be unwarranted. The question of how representative the large catch was of other potential catches in the area could have been dealt with at the time of the survey by adding more fishing sets in the stratum in an adaptive sampling type approach (Thompson 1992). The larger number of sets per stratum in the Georges Bank survey may have protected this survey from being dominated by one large set. More observations in stratum 56 of the eastern Scotian Shelf survey could have decreased the weighting associated with the large catch if such large catches were less common than one out of six sets. However, this survey was completed in 1988 and so the point is moot.

On the other hand, it may be quite useful to be able to postulate various alternatives for the distribution of catches in this stratum and evaluate their effect on the bootstrap confidence intervals. That is, given the observed catches (0, 1, 3, 106, 309, 5496), we may expect the distribution of possible catches to vary from a very smooth distribution over the range 0 to 5496 or higher, or to the very rough form implied by the survey where each catch has a one-sixth probability of occurring in the stratum. Four options that span this range of possible distributions are as follows: (*i*) the distribution of catches ranges continuously from 0 to some maximum above 5496 with a small probability of catches above 309; (*ii*) the distribution of catches represents a mixture of two groups, with the first five catches coming from the first group and the large catch of 5496 coming from the second; overlap in the range of catches from the two groups need not be expected and therefore there could be a zero probability for a range of catches somewhere between 309 and 5496; (*iii*) a more extreme case where the distribution of catches is made up of a number (less than six) of nonoverlapping distributions of catches; (*iv*) the distribution implied by fitting an empirical distribution function to the original data, resulting in each observation having a probability of one sixth assigned to it.

Distributions representing the first three options were fitted to the catches from stratum 56 using nonparametric density estimates. The fourth option was already used in constructing the original bootstrap estimates. For the first two options a Gaussian kernel was used with bandwidth estimated by the unbiased cross-validation estimator (Venables and Ripley 1994, p. 139). The distributions for the first option, estimated using an adaptive bandwidth estimate (Silverman 1986, pp. 101–102) and the second, which was estimated with a global bandwidth, are presented in Fig. 8. The original observations are indicated as X's on the abscissa. The distribution obtained with the global bandwidth assigned near-zero probabilities to all observations between 2100 and 4000 fish. The distribution for the third case (Fig. 9) was estimated with the discrete kernel estimate (Rajagopalan and Lall 1995) approach used for the Georges Bank data in the previous subsection. The large spread and sparseness of the observations resulted in four separate groups of discrete distributions when the discrete kernel estimate was used. When ordered from the original data, discrete kernel, global, and adaptive fit, these distributions represent increasing smoothness of the expected distribution of the numbers of haddock caught.

If these distributions are treated as predicting possible catches in the finite population sense, then the impact of these different hypothesized distributions on bootstrap confidence intervals can be evaluated in the following way. Using an approach similar to the smoothed bootstrap (Silverman 1986, p. 143), observations are generated randomly from the fitted distributions for the bootstrap resampling. To keep the observations positive, observations were generated directly from the cumulative density (mass) functions for the distributions in Figs. 8 and 9. A total of 1000 bootstrap replications were made for each of the three distributions for stratum 56 and the BWR method was used for the catches in the other strata.

Quantile–quantile plots of the 1000 bootstrap replications for each distribution are compared with that from using the BWR method (option 4) on all strata in Fig. 10. All of the cases have heavier tails than expected for a normal distribution. The distribution of replications from the distribution using the adaptive bandwidth was the smoothest of the four as expected. The other distributions became progressively rougher as the number of groups increased from two (global bandwidth) through to the original BWR estimates.

The bootstrap estimate of the variance decreased with

Fig. 8. Probability density functions estimated for the observed catches of haddock (indicated as X's on abscissa) from stratum 56 of the eastern Scotian Shelf survey (1988). Probability density functions were estimated using nonparametric density estimates with Gaussian kernels and unbiased cross-validation bandwidth estimates.

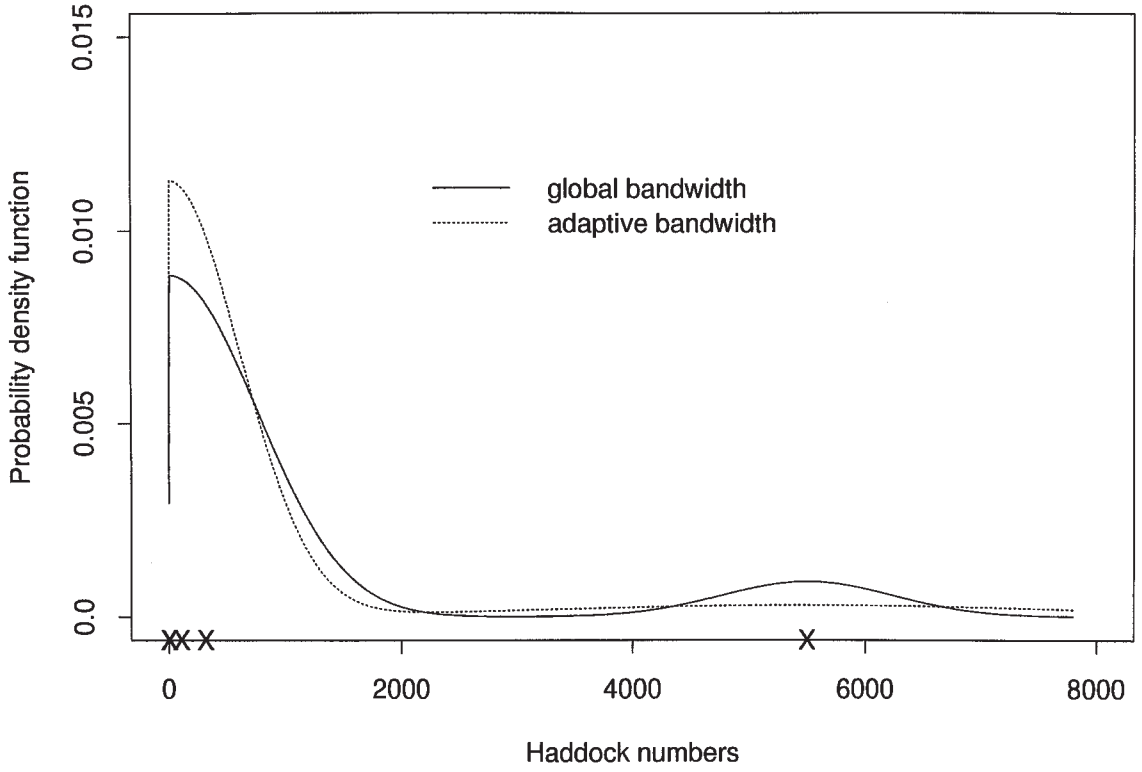


Fig. 9. Probability mass function estimated for the observed catches of haddock (indicated as X's on abscissa) from stratum 56 of the eastern Scotian Shelf survey (1988). Probability mass function was estimated using a discrete kernel estimate and a discrete bandwidth estimated by a cross-validation estimate.

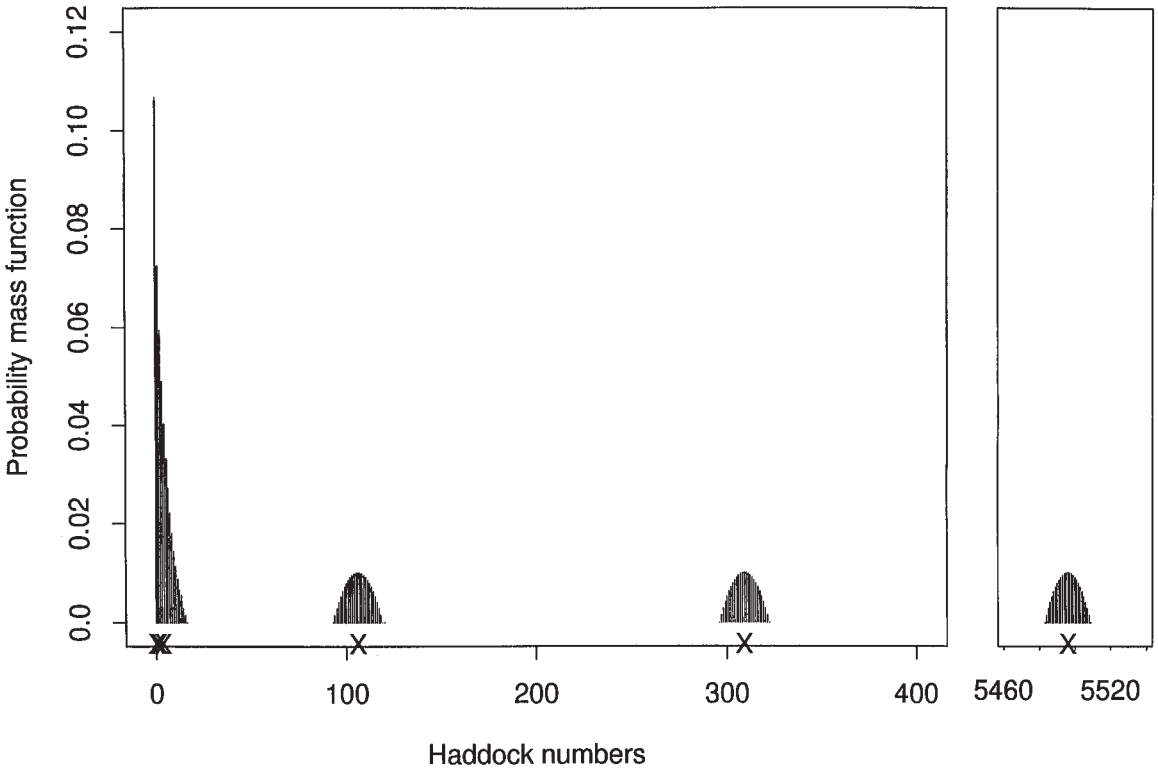


Fig. 10. Quantile–quantile plots showing the distributions of BWR bootstrap estimates of the stratified mean using all of the catches (standard) and using data simulated from each of the distributions in Figs. 8 and 9.

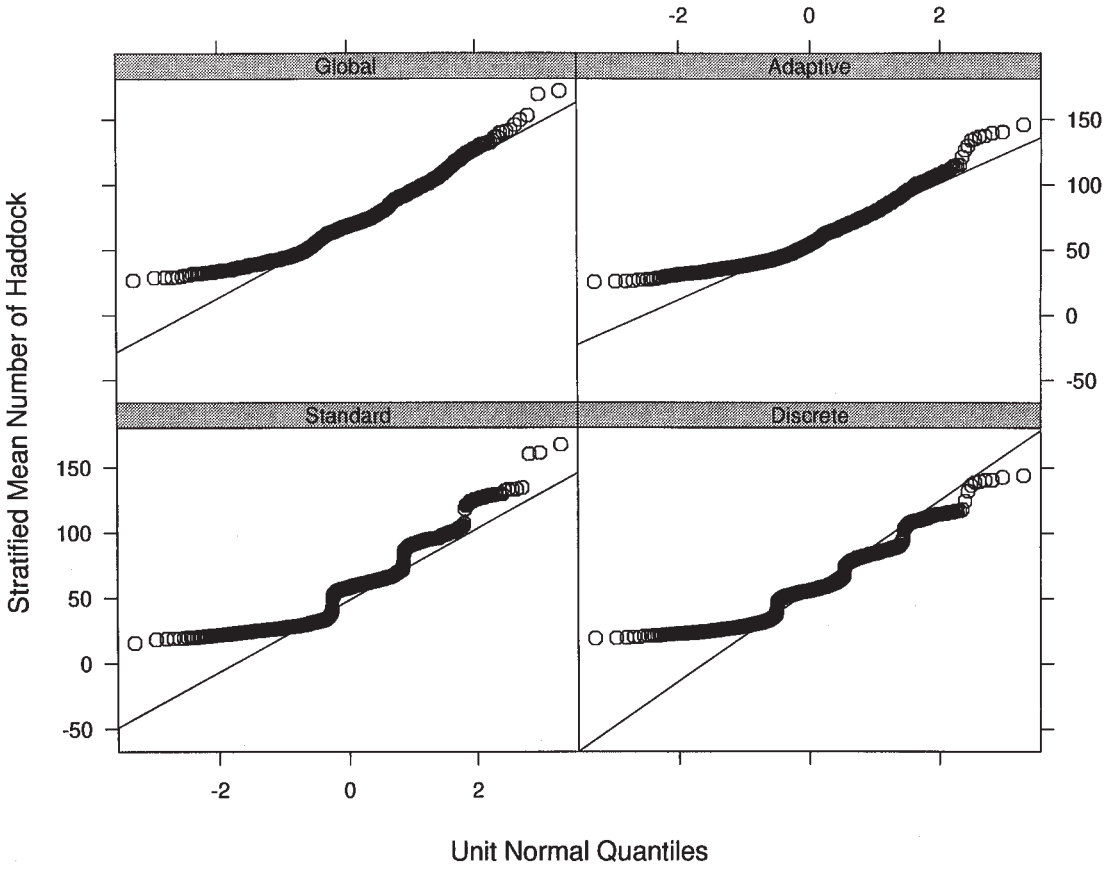


Table 5. Summary of results of the stratified mean, median, variance, and confidence intervals for the number of haddock caught in the 1988 eastern Scotian Shelf survey when nonparametric density estimates are used for the distribution of haddock catch in stratum 56.

				95% confidence interval				
Method	Mean	Median	Variance	Type	Lower	Upper	Length	Shape
Adaptive	58.68	54.22	447.1	PC	31.58	106.74	75.15	0.84
				BC	32.91	114.45	81.55	0.49
Global	69.84	68.33	594.2	PC	34.22	126.21	92.39	0.54
				BC	35.09	131.07	95.99	0.51
Discrete	58.07	55.66	667.3	PC	22.95	114.30	91.35	0.58
				BC	24.32	116.99	92.67	0.43

Note: Three methods are compared: the global bandwidth estimate and adaptive bandwidth estimate (both assuming a continuous distribution) and a Kernel estimate (assuming a discrete distribution). Confidence intervals for the bootstrap mean were calculated using the percentile (PC) and bias-corrected (BC) methods. The length column refers to the length of the confidence intervals. The shape column refers to a measure of symmetry (symmetric: shape = 0) given by Efron (1992).

increasing smoothness of the data (Table 5). The mean was the largest for the global option because of larger density associated with the wide range of the larger catches. The general trend for the PC limits was for narrower limits as the degree of smoothness increased. The upper limit for the PC method from the adaptive distribution was the lowest, but allowing for observations between 309 and 5496 fish (and above) increased the lower limit substantially. The lower limit for the global distribution was also higher than that from the original data but the upper limit was the closest to that from the original data (Table 3). Overall, the effect of assuming increasingly smoother distributions for haddock catches in stratum 56 on the resultant bootstrap PC intervals was to decrease the length and increase the lower limit. Also, comparing the upper PC limit from the adaptive distribution with those from the global and original data suggests that allowing for a continuum of catches over the range observed results in a much lower upper limit. The fitted bandwidth for the discrete distribution was quite narrow and the maximum catch was much lower than those for the global and adaptive distributions. Therefore, the upper PC limit was also lower than those from both the global distribution and the original data.

The behaviour of the BC limits[1] can be best explained by noting the changing relation between the mean and median for each distribution. For all of the smoothed bootstrap cases the median was less than the bootstrap estimate and hence upper and lower BC limits were greater than their respective PC limits. The reverse was true for the original BWR BC limits (Table 3).

Discussion

The results presented here illustrate that careful attention must be paid to incorporating the sampling scheme into the bootstrap method when dealing with complex designs. The naïve method would appear to be a straightforward extension of the bootstrap from the standard case to sampling with equal probability within each stratum. However, closer inspection shows that this method does not capture all of the aspects of the sampling scheme and an inconsistent estimate of the variance results. Getting the variance of the bootstrap estimates of the mean right is important because these estimates are used to derive the confidence limits.

The studies by Kimura and Balsinger (1985) and Sigler and Fujioka (1988) applied the bootstrap to estimate the abundance of sablefish (*Anoplopoma fimbria*) from complex survey designs that include strata defined by depth. The first study consisted of a pot survey conducted along the west coast of the United States and southeast coast of Alaska whereas the second was a longline survey in the Gulf of Alaska. In the first case, depth strata were referred to but not used in the calculation of abundance and therefore the bootstrap estimates were made assuming all observations had an equal chance of being chosen. However, in the longline survey, the area associated with the depth strata was used in estimating relative population numbers and the application of the bootstrap in this case was similar to the naïve bootstrap. The aim of the bootstrap analysis of the longline data was to detect changes in the estimated relative population numbers from year to year. Given that the naïve bootstrap underestimates the variance, it is probable that more annual differences in abundance were found for the longline survey than actually existed.

Fletcher and Webster (1996) produced results suggesting that a studentized version of the bootstrap augmented with an adjustment for skew will produce more accurate confidence intervals for the stratified mean than the bootstrap alone or the Student *t* method used here. However, it is difficult to compare their results with those given here because they used the naïve bootstrap and hence underestimated the variance.

The rescale and BWR (mirror-match) methods were designed to capture the second-order term of the Edgeworth expansions of the stratified mean with the bootstrap approximation (Chen and Sitter 1993) and therefore incorporate properties of the original estimates and the sampling scheme. The BWR method appears to be preferred to the rescale method from our results on the basis of simplicity and only being limited to a minimum sample size of $n_h = 2$. Sitter (1992*b*) compared the performance of the rescale methods (for $m_h = n_h - 1$ and $n_h - 3$) with various forms of the mirror-match methods (including BWR) for estimating confidence intervals for functions of stratified means (ratios, regression estimators, and correlation coefficients) and the stratified median. The mirror-match methods performed better on average than the rescale methods with respect to the accuracy of the respective confidence intervals being estimated. The results in Table 4 indicate that PC confidence limits from the BWR method were closest to those expected for the population of stratified means for the Georges Bank data.

All of the bootstrap methods gave insight into the suitability of the Student *t* distribution for constructing confidence intervals for the stratified mean. The fewer strata, larger sample sizes, and lower variability of the Georges Bank data resulted in the bootstrap limits that appeared to be similar to those given by the Student *t* distribution. However, the simulation results indicated that the small differences in the bootstrap limits were actually improvements over the Student *t* limits with respect to tail probabilities and coverage. In addition, these results also argued for the PC confidence interval method being quite adequate for constructing bootstrap confidence intervals for the BWR and rescale ($m_h = n_h - 3$) bootstrap methods.

For the eastern Scotian Shelf data, the relatively large number of strata, small sample sizes within strata, and higher variability resulted in very wide confidence intervals with a negative lower bound for the Student *t* method. In this case the bootstrap methods offered alternative intervals that were not as wide and resulted in positive lower limits. It was not possible to evaluate the performance of these confidence intervals given the small sample sizes within strata. While the BC_a method corrected the limits in the right direction for a skewed distribution, the simulation results in Table 4 suggest that this correction may be too severe.

All confidence interval methods were affected by the very large catch of haddock in stratum 56 of the Scotian Shelf survey. In practice, there seem to be four ways of dealing with large observations judged to be extreme: keep them, remove them, replace them, or reweight them. Keeping the large observations results in the current situation where we have a large variance, with one or a few catches driving the mean and confidence intervals so large as to be useless. Removing the observation outright is arbitrary and is probably not going to be very popular with the fishing industry: they may consider the one large catch as the only useful catch of the survey. Replacing the large catch(es) with the next largest catch (Smith 1981) or a function of the remaining observations (Moyer and Geissler 1991) could result in a smaller mean-square error but unknown bias.

Reweighting the sets in stratum 56 requires a basis for changing the current weighting of $1/n_h$. As noted before, the catch of 5496 was the largest catch of haddock in the entire time series of the survey to the present day. At 19 m this was the shallowest set made in stratum 56 and the near-bottom temperature of 10.98°C was the highest observed in the whole survey in 1988. In fact, sets have been made at 19 m in stratum 56 only twice in the history of the survey with the second such set made in 1990 resulting in a catch of 3271 haddock at a temperature of 7.36°C. Studies have shown that haddock appear to have strong positive associations with bottom temperature and depth when caught by trawl (Smith et al. 1994). It may be possible to determine the appropriate area weighting for each set on the basis of the expected distribution of depth and

[1] It was not obvious how to calculate the acceleration when smoothed data were used and therefore BC_a intervals were not calculated for these data

temperature in the stratum. The influence of the large set on the stratified mean and variance would probably be greatly decreased if only a small area of the stratum exhibited these apparently favourable conditions. The bootstrap procedure could be modified for this different weighting of the observations. However, while the depth contours of the stratum could be obtained to determine a depth–catch relationship, temperatures are only known for positions where the sets were made in any one year. A wider sampling of bottom conditions than this would be needed to objectively determine the distribution of temperature (or any other important hydrographic variable) in the stratum.

Are there other options for reweighting the set? When sampling fractions are small and extreme observations are few, Hidiroglou and Srinath (1981) suggest using the following weighting for t extreme observations,

$$\bar{y}_h^t = \frac{1}{N_h}\sum_{i=1}^{t} y_{hi} + \frac{N_h - t}{N_h(n_h - t)}\sum_{i=t+1}^{n_h} y_{hi}$$

based on comparing the ratio of its mean-squared to that of the standard estimator of the mean. Without any further guidance, assigning a weight of $1/N_h$ (= 1 / 80 928 for stratum 56) to the large catch may seem extreme and as arbitrary as simply removing the observation. Fitting a statistical distribution such as the Δ distribution (Pennington 1983) is another way of reweighting all of the observations in the stratum on the basis of an assumed mean–variance relationship. The nonparametric density (mass) estimates were used in the same spirit here but offered more flexibility with respect to the shape of the distribution. While six observations are really not enough to characterize a distribution very well, the resultant curves did allow for testing the effects of differently shaped distributions on the bootstrap confidence interval estimates. Use of nonparametric density estimation with bootstrap methods offers a very powerful combination of tools for investigating statistical issues where theoretical support is not well developed.

Acknowledgements

The data used in this report were diligently collected and processed by scientific staff of the Biological Station in St. Andrews, N.B., and the Marine Fish Division of the Bedford Institute of Oceanography. I also gratefully acknowledge the contributions of the officers and crew of the research vessel *Alfred Needler* to the successful operation of the research surveys. I thank Stratis Gavaris (Biological Station, St. Andrews) for his comments on an earlier draft of this paper. Jim Bence (Michigan State University), an anonymous referee, and associate editor Bill Warren provided many insightful and challenging comments on the submitted draft. Finally, I am grateful for the very valuable help on nonparametric density estimates provided by Balaji Rajagopalan (Columbia University, New York) and Peter Perkins (National Marine Fisheries Service, La Jolla, Calif.). Drs. Rajagopalan and Perkins were very generous to me with their software, time, and experience. The data sets used in this paper are available from the author via electronic mail for interested readers.

References

Azarovitz, T.R. 1981. A brief historical review of the Woods Hole trawl survey time series. *In* Bottom trawl surveys. *Edited by* W.G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. No. 58. pp. 62–67.

Buckland, S.T., Cattanach, K.L., and Anganuzzi, A.A. 1992. Estimating trends in abundance of dolphins associated with tuna in the eastern Tropical Pacific Ocean, using sightings data collected on commercial tuna vessels. Fish. Bull. **90**: 1–12.

Chambers, R.L., and Dunstan, R. 1986. Estimating distribution functions from survey data. Biometrika, **73**: 597–604.

Chen, J., and Sitter, R.R. 1993. Edgeworth expansion and the bootstrap for stratified sampling without replacement from a finite population. Can. J. Stat. **21**: 347–357.

Cochran, W.G. 1977. Sampling techniques. John Wiley & Sons, Inc., New York.

Doubleday, W.G. (*Editor*). 1981. Manual on groundfish surveys in the Northwest Atlantic. NAFO Sci. Counc. Stud. No. 2.

Efron, B. 1981. Nonparametric standard errors and confidence intervals. Can. J. Stat. **9**: 139–172.

Efron, B. 1982. The jackknife, the bootstrap and other resampling plans. Vol. 38. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM.

Efron, B.1992. Jackknife-after-bootstrap standard errors and influence functions (with discussion). J. R. Stat. Soc. Ser. B, **54**: 83–127.

Efron, B., and Tibshirani, R.J. 1993. An introduction to the bootstrap. Chapman & Hall, London.

Fletcher, D., and Webster, R. 1996. Skewness-adjusted confidence intervals in stratified biological surveys. J. Agric. Biol. Environ. Stat. **1**: 120–130.

Gavaris, S., and Eeckhaute, L.V. 1990. Assessment of haddock on eastern Georges Bank. CAFSAC research document No. 90/86. Canadian Atlantic Fisheries Scientific Advisory Committee, Dartmouth, N.S.

Gunderson, D.R. 1993. Surveys of fisheries resources. John Wiley & Sons, Inc., New York.

Halliday, R.G., and Koeller, P.A.1981. A history of Canadian groundfish trawling surveys and data usage in ICNAF divisions 4TVWX. *In* Bottom trawl surveys. *Edited by* W.G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. No. 58. pp. 27–42.

Hidiroglou, M.A., and Srinath, K.P. 1981. Some estimators of a population total from simple random samples containing large units. J. Am. Stat. Assoc. **76**: 690–695.

Jolly, G.M., and Smith, S.J. 1989. A note on the analysis of marine survey data. *In* Progress in fisheries acoustics: proceedings of the Institute of Acoustics, MAFF Fisheries Laboratory, Lowestoft, England. Vol. 11. Part 3. Institute of Acoustics, MAFF Fisheries Laboratory, Lowestoft, England. pp. 195–201.

Jones, M.C., and Bradbury, I.S. 1993. Kernel smoothing for finite populations. Stat. Comput. **3**: 45–50.

Kimura, D.K., and Balsiger, J.W. 1985. Bootstrap methods for evaluating sablefish pot index surveys. N. Am. J. Fish. Manage. **5**: 47–56.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. 1988. Bootstrap and other methods to measure errors in survey estimates, Can. J. Stat. **16**(Suppl.): 25–45.

McConnaughey, R.A., and Conquest, L.L. 1993. Trawl survey estimation using a comparative approach based on lognormal theory. Fish. Bull. **91**: 107–118.

Moyer, L.M., and Geissler, P.H. 1991. Accomodating outliers in wildlife surveys. Wildl. Soc. Bull. **19**: 267–270.

Myers, R.A., and Pepin, P. 1990. The robustness of lognormal-based estimators of abundance. Biometrics, **46**: 1185–1192.

Pelletier, D., and Gros, P. 1991. Assessing the impact of sampling error on model-based management advice: comparison of equilibrium yield

per recruit variance estimators. Can. J. Fish. Aquat. Sci. **48**: 2129–2139.
Pennington, M. 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics, **39**: 281–286.
Pitt, T.K., Wells, R., and McKone, W.D. 1981. A critique of research vessel otter trawl surveys by the St. John's Research and Resource Services. *In* Bottom trawl surveys. *Edited by* W.G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. No. 58. pp. 42–61.
Rajagopalan, B., and Lall, U. 1995. A kernel estimator for discrete distributions. Nonparam. Stat. **4**: 409–426.
Rao, J.N.K., and Wu, C.F.J. 1988. Resampling inference with complex survey data. J. Am. Stat. Assoc. **83**: 231–241.
Robotham, H., and Castillo, J. 1990. The bootstrap method: an alternative for estimating confidence intervals of resources surveyed by hydroacoustic techniques. Rapp. P.V. Reun. Cons. Int. Explor. Mer, **189**: 421–424.
Scott, D.W. 1992. Multivariate density estimation: theory, practice and visualization. John Wiley & Sons, Inc., New York.
Sigler, M.F., and Fujioka, J.T. 1988. Evaluation of variability in sablefish, *Anoplopoma fimbria,* abundance indices in the Gulf of Alaska using the bootstrap method. Fish. Bull. **86**: 445–452.
Silverman, B.W. 1986. Density estimation for statistics and data analysis. Chapman and Hall, London.
Sitter, R.R. 1992*a*. A resampling procedure for complex survey data. J. Am. Stat. Assoc. **87**: 755–765.
Sitter, R.R. 1992*b*. Comparing three bootstrap methods for survey data. Can. J. Stat. **20**: 135–154.
Skinner, C.J., Holt, D., and Smith, T. (*Editors*). 1989. Analysis of complex surveys. John Wiley & Sons, Inc., New York.
Smith, S.J. 1981. A comparison of estimators of location for skewed populations, with application to groundfish trawl surveys. *In* Bottom trawl surveys. *Edited by* W.G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. No. 58. pp. 154–163.
Smith, S.J. 1988*a*. Abundance indices from research survey data. *In* Collected papers on stock assessment methods. *Edited by* D. Rivard. CAFSAC research document No. 88/61. Canadian Atlantic Fisheries Scientific Advisory Committee, Dartmouth, N.S. pp. 16–43.
Smith, S.J. 1988*b*. Evaluating the efficiency of the Δ-distribution mean estimator. Biometrics, **44**: 485–493.
Smith, S.J. 1990. Use of statistical models for the estimation of abundance from groundfish trawl surveys. Can. J. Fish. Aquat. Sci. **47**: 894–903.
Smith, S.J., and Gavaris, S. 1993. Evaluating the accuracy of projected catch estimates from sequential population analysis and trawl survey abundance estimates. *In* Risk evaluation and biological reference points for fisheries management. *Edited by* S.J. Smith, J.J. Hunt, and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. No. 120. pp. 163–172.
Smith, S.J., Losier, R.L., Page, F.H., and Hatt, K. 1994. Associations between haddock, and temperature, salinity and depth within the Canadian groundfish bottom trawl surveys (1970–1993) conducted in NAFO divisions 4VWX and 5Z. Can. Tech. Rep. Fish. Aquat. Sci. No. 1959.
Stanley, R.D. 1992. Bootstrap calculation of catch-per-unit-effort variance from trawl logbooks: do fisheries generate enough observations for stock assessments? N. Am. J. Fish. Manage. **12**: 19–27.
Taylor, C.C. 1953. Nature of variability in trawl catches. Fish. Bull. **54**: 145–166.
Thompson, S.K. 1992. Sampling. John Wiley & Sons, Inc., New York.
Venables, W., and Ripley, B. 1994. Modern applied statistics with S-Plus. Springer-Verlag, New York.

Appendix

Stratified estimate of acceleration

Efron and Tibshirani (1993) define the acceleration a as the rate of change of the standard error on the normalized scale. For one-parameter models, they offer the following as a good approximation for a:

$$\hat{a} = \frac{1}{6}\text{skew}_{\theta=\hat{\theta}}(l_\theta),$$

where (l_θ) is the score function. An estimate for this quantity is (Efron and Tibshirani 1993, p. 186)

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\left(\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(I)})^2\right)^{3/2}}$$

The $\hat{\theta}_i$ refer to the jackknife values of the statistic $\hat{\theta}$ and $\hat{\theta}_\cdot = \Sigma_{i=1}^{n}\hat{\theta}_i/n$. Note that the denominator and numerator are similar to the second (with power 3/2) and third central moments of $\hat{\theta}$ in the standard definition of skew — μ_3/σ^3. The second central moment for the stratified mean is the variance given in eq. 2, while the third central moment is defined as (Rao and Wu 1988, p. 326)

$$\hat{\mu}_3(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^3(1-f_h)(1-2f_h)\hat{\mu}_{3h}$$

where

$$\hat{\mu}_{3h} = \sum_{i=1}^{n_h}\frac{(y_{hi}-\bar{y}_h)^3}{n_h(n_h-1)(n_h-2)}$$

Replacing s_h^2/n_h and $\hat{\mu}_{3h}$ with the above jackknife quantities for each stratum gives the following as an estimate of the acceleration for a stratified random design:

$$\hat{a}_{st} = \frac{\sum_{h=1}^{L} W_h^3(1-f_h)(1-2f_h)\sum_{i=1}^{n_h}(\hat{\theta}_{h(\cdot)} - \hat{\theta}_{h(i)})^3}{6\left(\sum_{h=1}^{L} N_h(N_h-n_h)\sum_{h=1}^{n_h}(\hat{\theta}_{h(\cdot)} - \hat{\theta}_{h(i)})^2/N^2\right)^{3/2}}$$

In this paper, jackknife values for the stratum means are substituted for $\hat{\theta}_{h(i)}$.